Operationalizing the Fundamental Rights Impact Assessment for AI systems: the FRIA Project

Roberta Savella[1,*], Francesca Pratesi[1,*], Roberto Trasarti[1], Lucilla Gatt[2], Maria Cristina Gaeta[2], Ilaria Amelia Caggiano[2], Livia Aulino[2], Emiliano Troisi[2] and Luigi Izzo[2]

[1]*Institute of Information Science and Technologies, National Research Council of Italy, via Moruzzi 1, Pisa, 56124, Italy*

[2]*Research Centre of European Private Law (ReCEPL), Università degli Studi Suor Orsola Benincasa, C.so Vittorio Emanuele, 334, 80135, Naples, Italy*

Abstract

This paper presents the FRIA Project, a multidisciplinary research study which connects the legal and ethical aspects related to the impact on fundamental rights of Artificial Intelligence systems and the technical issues that arise in the creation of an automated tool for the Fundamental Rights Impact Assessment, which is the ultimate objective of this work.

Keywords

Fundamental Rights Impact Assessment, AI Act, Automated assessment, High-risk AI systems

1. Introduction

At the time of writing of this paper, the European Union institutions are finalizing the adoption of the world's first comprehensive law on Artificial Intelligence (AI): the AI Act [1]. The final text maintains the risk-based approach proposed since 2021 by the Commission, with different rules applicable to AI systems based on the level of risk they pose. While the technologies which create an "unacceptable risk" are banned, the "high-risk" category is heavily regulated, with significant requirements for the systems and obligations for the providers, importers, distributors, and deployers of these technologies. However, the first version of the text proposed in 2021 was significantly amended to take into account the technological progress in this field (for example, the development of general-purpose AI) and the instances of institutions, advocates and associations. One of the most significant changes follows the calls of several organizations and scholars [2, 3] and introduces the obligation for some deployers - bodies governed by public law, or private operators providing public services, or operators deploying high-risk systems that evaluate the creditworthiness of natural persons, establish their credit score, or use AI for risk assessment and pricing in the case of life and health insurance - to carry out a Fundamental Rights Impact Assessment ("FRIA"). The FRIA is required when the aforementioned deployers put into use certain high-risk systems, as listed in Annex III of the AI Act [1], with the exception of the systems used for critical infrastructures. Article 27 of the final text voted by the European Parliament the 13th of March 2024 [1] identifies the cases in which the FRIA is mandatory and the content of the assessment, but it does not provide indications on parameters and criteria for the implementation of adequate measurement paths. Therefore, the project we are presenting with this paper starts from the identification of the need for a practical solution to operationalize the legal requirement of the FRIA. In this paper, we would like to present a methodology to operationalize the FRIA requirements, and make a first step in the automatization of the assessment related to the fundamental rights.

2. Methodology

The FRIA Project adopts a multidisciplinary approach to connect the legal and ethical study of how Artificial Intelligence systems can impact on fundamental rights and the technical aspects related to the creation of an automated impact assessment tool. The project will be articulated in four phases, described in the following subsections.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

[*] Corresponding author.

[†] These authors contributed equally.

roberta.savella@isti.cnr.it (R. Savella); francesca.pratesi@isti.cnr.it (F. Pratesi); roberto.trasarti@isti.cnr.it (R. Trasarti); lucilla.gatt@unisob.na.it (L. Gatt); mariacristina.gaeta@unisob.na.it (M. C. Gaeta); ilaria.caggiano@unisob.na.it (I. A. Caggiano); livia.aulino@gmail.com (L. Aulino); emilianotroisi@gmail.com (E. Troisi); gigizzom@gmail.com (L. Izzo)

0009-0007-3005-4252 (R. Savella); 0000-0002-4260-4473 (F. Pratesi); 0000-0001-5316-6475 (R. Trasarti); 0000-0001-7730-7180 (L. Gatt); 0000-0001-5789-2079 (M. C. Gaeta); 0000-0001-7539-3695 (I. A. Caggiano); 0000-0003-2435-4834 (L. Aulino); 0000-0002-0560-0916 (E. Troisi); 0000-0002-6826-7452 (L. Izzo)

2.1. Assessment of the legal and ethical requirements for AI to ensure the respect of fundamental rights

In this first phase, the research will be focused on the analysis of the current legal and ethical framework regarding Artificial Intelligence and fundamental rights, starting from the final version of the AI Act, the Universal Declaration of Human Rights [4], the European Convention of Human Rights [5], the Charter of Fundamental Rights of the European Union [6], but also the UN Sustainable Development Goals [7], and all the relevant academic research, guidelines and best practices concerning fundamental rights and human rights impact assessment. The research team will also take into account pre-existing tools used for impact assessment of new technologies, such as, for example, the Assessment List for Trustworthy Artificial Intelligence (ALTAI) [8]. The expected outcome of this step is to identify specific requirements for AI systems regarding fundamental rights protection and sustainability.

2.2. Identification of parameters to ensure compliance with the obligations

In this phase, we will determine technical parameters and standards to translate the requirements identified in the first phase into practical and quantifiable indicators and requisites. This will be done taking into account also pre-existing risk assessment evaluation frameworks and standards applicable to FRIA. Using the identified parameters, a theoretical methodology for the FRIA of AI systems will be developed. The methodology will make it possible to identify potential non-compliances (or gaps) with the regulatory requirements. The methodology will serve as a basis and input for the development of a tool prototype. This phase will be the most relevant part of the project, and we expect to have a significant impact on the current landscape of assessment frameworks and criteria, providing an innovative methodology to take into account all the relevant aspects related to the impact of new technologies on fundamental rights and sustainability in the short, medium and long term.

2.3. Translation of the parameters into a prototype to support and automate the Fundamental Rights Impact Assessment of AI systems

In this phase, the researchers will develop the prototype to support and automate the FRIA for some of the metrics which will be selected due to feasibility with regard to the state of the art in this field. The prototype will be designed to be easy-to-use and understandable by its users. First, we will identify an evaluation metric that will be completely open and as objective as possible, to be sure that all the final users have a clear indication on how to answer every question. Then, we need to assign a score to each answer, in order to provide an objective evaluation of each dimension. At the end of the process, the prototype will provide a global score of the system, a set of scores highlighting the strong and the weak points of the AI system, based on the obtained information. These scores will be presented in a succinct graphical form, e.g., throughout a radar plot (see Figure 1 for an example), where each dimension represents one of the relevant fundamental rights.

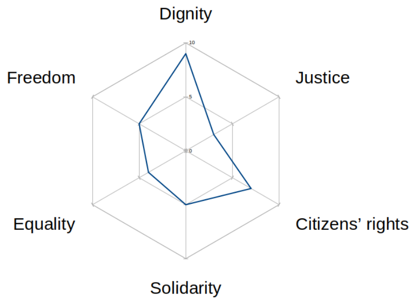

Figure 1: Example of the outcome summarizing the scores obtained by an AI system among the various fundamental rights, reported in the axis.

In particular during this phase the research activity will focus on the AI assistance systems used in the judicial field as a relevant area of application cited in Annex III of the AI Act [1]. This will give us the possibility of focusing on a specific context and giving us a test base for our platform.

2.4. Indicators, requisites and prototype validation

In this last part of the project, the indicators, requisites, and prototype will be tested and validated through focus groups. The validation will be carried out in two phases of the project, the first one before implementation of the prototype and the second one at the end of the project. This activity will also include a case study in which the validation process of the prototype will be applied to an existing AI technology and in particular on the administration of justice systems. In order to carry out this phase, we will also take advantage of the 'regulatory sandboxes' (regulated in Art. 57-63 of the AI Act [1]), i.e., a mechanism established to foster innovation in AI, experimenting and testing in a controlled environment new products and services under a regulator's supervision.

3. Conclusions

In this paper we presented the line of work our research group intends to follow in order to develop an automated tool to operationalize the FRIA. It is important to point out that this study is based on the synergic interaction between legal and IT professionals, with the objective of embodying the abstract legal and ethical principles and obligations into a technological solution. For this reason, one of the most challenging efforts in our research will be to translate the requirements for the FRIA into quantifiable and machine-readable metrics. Moreover, as the field of AI is still developing, with new technologies and new regulations emerging at an incredible pace, another critical point will be closely monitoring the state of the art to select during the project a specific area of application to design the tool.

Acknowledgments

This work is partially supported by the European Union – Horizon 2020 Program under the scheme “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities”, Grant Agreement n.871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” (http://www.sobigdata.eu) and by the Research Centre of European Private Law (ReCEPL), coordinated by the ReCEPL Director Prof. Lucilla Gatt.

References

[1] Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), P9_TA(2024)0138, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024. Last Accessed: 2024-04-23.

[2] Urgent appeal to approve a solid Fundamental Rights Impact Assessment in the EU Artificial Intelligence Act, https://brusselsprivacyhub.com/2023/09/12/brussels-privacy-hub-and-other-academic-institutions-ask-to-approve-a-fundamental-rights-impact-assessment-in-the-eu-artificial-int 2023. Last Accessed: 2024-04-23.

[3] The AI Act Must Protect the Rule of Law, https://dq4n3btxmr8c9.cloudfront.net/files/iytbh9/AI_and_RoL_Open_Letter_final_27092023.pdf, 2023. Last Accessed: 2024-04-23.

[4] Universal Declaration of Human Rights of the United Nation (UDHR) of 10 December 1948, https://www.un.org/en/about-us/universal-declaration-of-human-rights, 1948. Last Accessed: 2024-04-23.

[5] European Convention on Human Rights (ECHR) of 4 November 1950, https://www.coe.int/en/web/human-rights-convention/the-convention-in-1950, 1950. Last Accessed: 2024-04-23.

[6] Charter of Fundamental Rights of the European Union of 18 December 2000 (FREU), https://www.europarl.europa.eu/charter/pdf/text_en.pdf, 2000. Last Accessed: 2024-04-23.

[7] United Nations Sustainable Development Goals, https://sdgs.un.org/goals, 2023. Last Accessed: 2024-04-23.